DOCUMENT RESUME

ED 094 277 CG 008 978

AUTHOR Hawkins, Robert P.; Dotson, Victor A.
TITLE Reliability Scores That Delude: An Alice in Wonderland Trip Through the Misleading Characteristics of Inter-Observer Agreement Scores in Interval Recording.
PUB DATE [73]
NOTE 25p.

EDRS PRICE MF-$0.75 HC-$1.85 PLUS POSTAGE
DESCRIPTORS Behavioral Science Research; Evaluation Criteria; *Experiments; *Intervals; Observation; *Psychology; *Reliability; Research Design; Research Projects; *Response Style (Tests)
IDENTIFIERS *Interval by Interval Method

ABSTRACT

The purpose of this project was to assess the degree to which the most popular method of calculating reliability of interval data, the I-I method, served the author's scientific purposes. Data were collected and analyses performed to assess the adequacy of I-I reliability scores in serving three functions: (1) as an index of how precise, clear, objective and complete the definition is; (2) as an index of how competently the observer is recording; and (3) as an index of the believability of the experimental effect reported. Findings from the study indicate gross unreliability of I-I scores as an index of definition adequacy, observer competence or believability of experimental effects. This unreliability is attributed to the fact that I-I scores are highly subject to influence by the rate or duration of the behavior being recorded. Conclusions by the authors stress that, because the I-I reliability scores are clearly inadequate, it is likely that a significant body of applied behavior analysis has seriously misrepresented the relationships between certain environmental factors and certain significant human behaviors. (Author/PC)

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED 094277

Reliability Scores That Delude: An Alice in Wonderland Trip Through the Misleading Characteristics of Inter-Observer Agreement Scores in Interval Recording

Robert P. Hawkins and Victor A. Dotson
Western Michigan University

Bijou, Peterson and Ault (1968), in describing methods for recording data in the natural environment, give considerable emphasis to the interval method of recording. In the interval method, the experimental session is divided into equal time intervals for recording purposes, usually 10-second or 20-second intervals, but the size can be adjusted to suit the frequency or duration of the behavior or to make recording more convenient. For each time interval during the session the observer records whether the response was or was not occurring at any time during the interval (or, occasionally, for a certain portion of an interval). While this method has certain limitations, it also offers some outstanding advantages. It allows an observer to easily measure several responses concurrently (e.g., Hart, Reynolds, Baer, Brawley, and Harris, 1968; Madsen, Becker, and Thomas, 1968), it shows changes in either the frequency or the duration of a behavior, and it circumvents the sometimes difficult task of defining and detecting single units of behavior (as in the case where the experimenter wishes to record talking, cooperative play, or attending to a task).

Bijou, et al. (1968) also point out the necessity of measuring the inter-observer reliability of data recorded by observers (as opposed to data recorded automatically by mechanical, electromechanical, or electronic apparatus). The method they describe for calculating the agreement between the data recorded by two independent observers employs the following formula:

$$\frac{\text{agreements}}{\text{agreements + disagreements}} \times 100 = \%\ \text{agreement}$$

An agreement is any interval in which both observers recorded that the response was occurring during the interval or in which both observers recorded that the response did not occur during the interval. Disagreements are intervals in which only one observer reported that the response occurred. Thus, every interval of recording is used in the calculation of inter-observer agreement by this method. We shall call this method of calculation the "interval-by-interval" (I-I) method.

Of approximately 97 studies involving behavior analysis in education reported in the first five volumes of the Journal of Applied Behavior Analysis (1968-1972), approximately 40%, used interval recording for at least a portion of the data. Of these studies using interval recording, 70%, or 26, appear to have calculated inter-observer agreement by the formula given above.[1,2] Thus it clearly is the most popular method of calculating agreement (three other methods will be presented and discussed below). But how well does it serve our purposes as scientists?

In general, the purpose of checking the reliability of observer-recorded data is to assess the accuracy and objectivity of the data. But this is an insufficient analysis. First, there are at least 3 somewhat independent sources of error in obtaining accurate and objective data: (1) the definition of behavior given the observer by the experimenter may be vague, subjective or incomplete; (2) the observer may be poorly trained, unmotivated, or otherwise incompetent; and (3) the behavior may be difficult to detect because of its subtlety or complexity, because of distractions, or because of other factors obstructing the observing process. Second, it is not simply the accuracy and

objectivity of the data themselves that needs to be assessed, but also the "believability" or validity of the experimental effect reported (referred to by Campbell and Stanley, 1966, as internal validity). Of course when the data are perfectly accurate, the validity of experimental effect is also perfect; but the validity of experimental effect declines more rapidly than does the validity of any particular datum, when the validity of that datum is less than perfect, as will be shown later in this paper.

The purpose of the present research is to assess the degree to which the most popular method of calculating reliability of interval data, the I-I method, serves our scientific purposes. We collected data and performed other analyses to assess the adequacy of I-I reliability scores in serving three functions: (1) as an index of how precise, clear, objective and complete the definition is; (2) as an index of how competently the observer is recording; and (3) as an index of the believability of the experimental effect reported. The analyses of these three functions of reliability scores will be presented in that order.

The Adequacy-of-Definition Function

Test #1

In order to test the ability of I-I reliability scores to reflect the adequacy of response definitions we first wrote a few definitions that we felt were obviously absurd (that is, their face validity was low) in terms of completeness, precision, objectivity, and unambiguousness. These definitions were as follows:

Positive affect: any time the subject shows such feelings as pleasure, happiness, joy, affection, admiration or excited animation.

Negative affect: any time the subject is exhibiting such feelings as displeasure, anger, embarrassment, hate, disillusionment, discomfort, disappointment, fear, anxiety or sorrow.

Neutral affect: any time the subject is exhibiting neither negative nor positive affect, as defined above.[3]

The observers were also instructed that at least one of these affective states would necessarily be recorded in each interval, but that more than one could also occur within the same interval.

Two trained, experienced, graduate student observers were given written copies of the definitions and asked to study them for five minutes.[4] They then independently recorded the affect of a school child for 20 minutes, employing 10-second interval recording. Independence of recording was assured by erecting a cardboard barrier between the observers to prevent their seeing when or what the other observer recorded. Attentiveness to the task was assured by instructing the observers as to the importance of the data and having one of the authors sit where he could watch both observers. The observers were given no hint as to the authors' hypothesis or even that the authors considered the definitions inadequate. The I-I agreement between the observers on the first and only session was 95%, 92%, and 100% respectively for these behaviors.

Using the same method, the same two observers recorded each of the following behaviors for 20 minute sessions, employing the definitions indicated:

Thinking: any time the child appears to be considering something, weighing opposing impulses, arriving at the solution to a problem, etc.

Excessive movement: Any time the subject is showing more bodily motion than the situation calls for. The motion may be in his whole body, as in jumping up and down, or may be in just a small portion of his body, as in finger tapping.

Interest: Any time the subject shows by what he says, the expression on his face, the vigor of his movement, the intonation of his voice, or other aspects of his behavior that he is interested in a particular thing. He may show that he is interested in what someone is saying or doing, in his work, or in some play activity he is engaged in, or anything else.

Acting silly: Score an interval "S" any time that the child is acting below his age.

The I-I agreement scores on these four behaviors were 83%, 57%, 84% and 99% respectively. Only the score on "excessive" movement falls below acceptable levels. The reason for such high agreement scores despite the obvious inadequacy of the definitions will be discussed later in this paper.

Test #2

What if a definition were so ambiguous that two different observers presented with it drew completely different meanings from it? We obtained a test of this question by accident. Two independent observers were recording several behaviors in a classroom and one of them was writing the wrong symbols for two of the behaviors; he had reversed the symbols for writing and hand-raising on four successive sessions before the error was corrected. This situation can be considered a test of the degree to which I-I scores reflect definition adequacy in the sense that it is the same problem that might have occurred if two observers were simply told by an experimenter, "Now I'd like you to record this student's handraising behavior," without the behavior's being defined at all (except for its name), and if one of the observers thought the experimenter had said "handwriting" instead of "handraising." These behaviors, as defined, were topographically incompatible; the only way they could both be emitted at the same time would be for a person to be writing with one hand and raising his other hand, a highly unlikely combination of

responses in a junior high school classroom (perhaps somewhat more likely in a college class, where notes might be taken while waiting to contribute to discussion).

On those four sessions, in which one observer's data on writing were checked against the other observer's data on handraising, the observers attained agreement scores of 77%, 80%, 89%, and 92%. That is, even though they were recording different responses that rarely would (in fact, rarely did) occur in the same interval, they obtained high I-I reliability scores!

Clearly, I-I reliability scores must be considered highly insensitive as an index of the adequacy of response definitions. This is a scientifically significant limitation because the definition of a response is a very important aspect of the measurement process, and measurement is the first requirement of a science. If the adequacy of a definition cannot be assessed it is difficult for one scientist to know whether another scientist's definition will be useful in his work or even what behavior is really measured by that definition. In addition, observer biases are likely to be more influential when explicit definitions are inadequate. The reasons for the insensitivity of I-I reliability as an index of the adequacy of a definition will be discussed later.

The Observer-Competency Function

Perhaps the most extreme form of observer imcompetence would be to fall asleep. This is not unheard of, and is most likely to occur when a low-rate behavior is being recorded over a long sessions (say 45 minutes or more). To test I-I reliability as an index of observer competence, we simulated the condition where an observer falls asleep. We took the classroom data recorded

by a single observer in another experiment on three behaviors over four consecutive sessions. The behavior were teacher talking, student writing, and student handraising. These four data sheets (each with data on three behaviors) were compared with a blank sheet representing the data of a sleeping observer (who, of course, never saw the behavior occur during the session), and I-I reliability scores were calculated between each real data sheet and the blank one.[5] The agreement scores on teacher talking ranged for 2% to 54% (mean, 17%). These would have been reassuring scores except that the scores on handraising ranged from 86% to 98% (mean, 93%) and those on writing ranged from 77% to 100% (mean, 91%), despite the fact that one "observer" was asleep! Apparently I-I reliability scores provide a less adequate measure of observer competence than one might hope for.

The Believability-of-Experimental-Effect Function

If the experimental effects reported in the behavior analysis (or any other scientific) literature cannot be relied on as representing real phenomena, many of our research and programming activities could be wasted effort. To what extent do I-I reliability scores provide a safeguard against the reporting of effects that are not real or the failure to see effects that did occur? This question will first be examined by an analysis of a fictional behavior change study.

Suppose a child was a social isolate and spent a relatively small amount of time interacting with peers during times when such interaction was considered desirable. A behavior analyst comes to the rescue, recommending that the teacher apply some particular technique, perhaps one that happens to have been invented by this behavior analyst. An observer, perhaps the teacher, records baseline data daily by a 10 second interval method, finding that the behavior consistently fails to occur at all. Then the special technique is applied

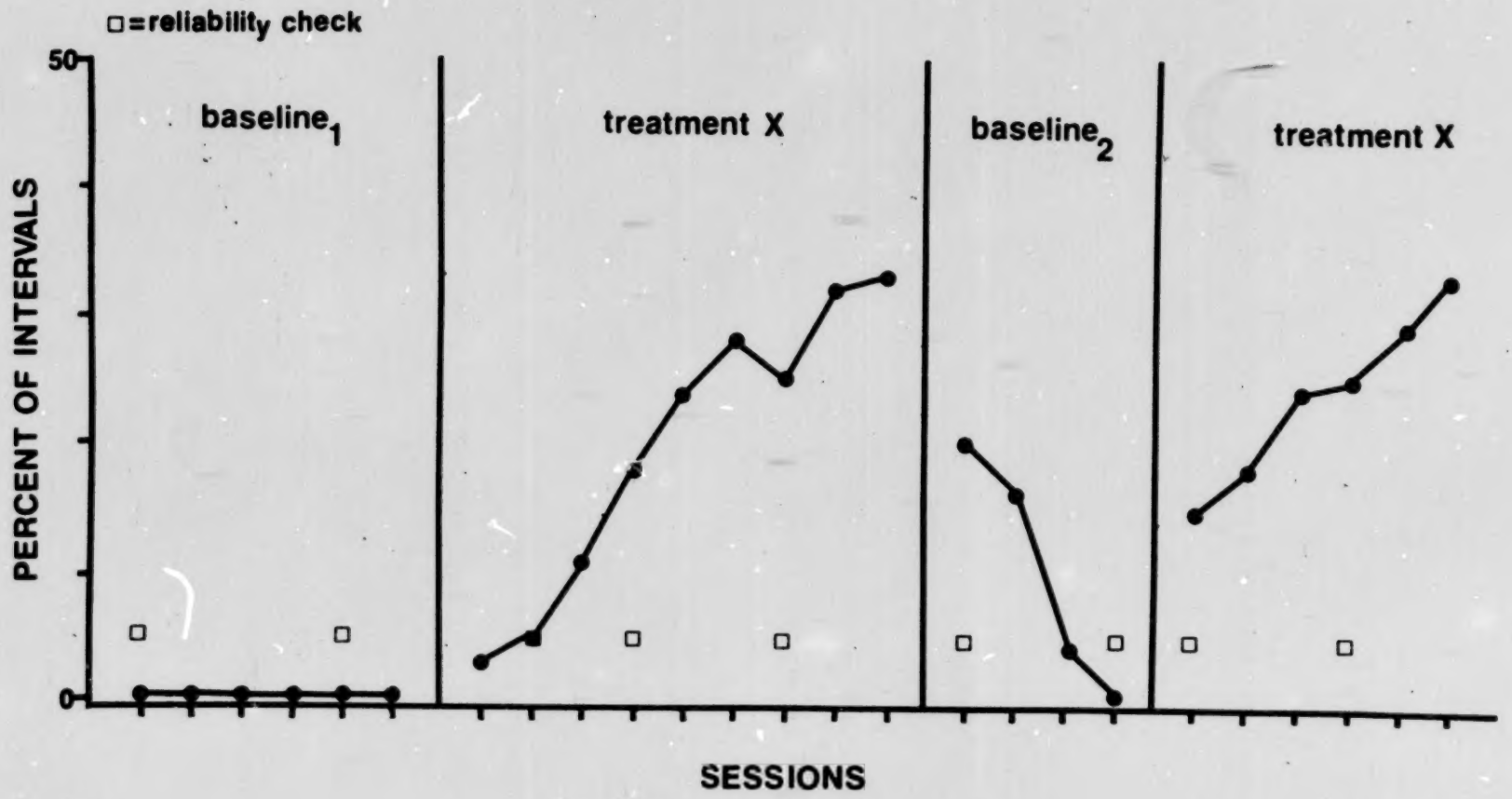
and the data show an increase until the behavior is reported to occur in approximately 30% of the intervals, which might be well within the norms in this group of children. Perhaps the effect is even replicated in a reversal design. These fictitious results are shown in Figure 1. There appears to be a very significant effect from the intervention.

Insert Fig. 1 about here.

But suppose that unknown to the first observer a second observer independently records that same behavior on several occassions, and that the second observer is totally objective, and accurate, perhaps because he or she was naive as to the nature of the experiment or the behavior analyst's personal interest in the success of the technique. And, further, suppose that this objective observer saw the behavior in exactly 5% of the intervals on every reliability check; that is, the behavior showed no change whatever during the study (represented by the squares in Figure 1). If the second observer happened to record on sessions 1, 5, 8, 10, 13, 16, 19, 20, and 23, the I-I agreement scores could have had a range as high as 80% to 100%, with a mean of 90%; and the lowest possible range of I-I agreement scores would have been from 70% to 95% with a mean of 83%! Thus, even though reliability was checked relatively, I-I reliability scores could not warn the experimenter (or the reader of his report) that the data was grossly biased toward finding an experimental effect.[6]

This raises a serious question as to the ability of I-I scores to protect against significant distortion of experimental effects. However, examination of some published data would provide a more adequate impression of the practical implications of this potential distortion. We selected a classic study to examine: the first study in the first paper in the first issue of the Journal of Applied Behavior Analysis, a study reported in the frequently-cited paper by Hall, Lund and Jackson (1968).[7]

The data from this study, dealing with the studying behavior of a child named Robbie, were obtained by interval recording; and the interval-by-interval method was employed to calculate inter-observer agreement. The paper does not report how many reliability checks were made, but simply


□=reliability check
50
0
PERCENT OF INTERVALS
baseline1
treatment X
baseline2
treatment X
SESSIONS

that there were several and that they included at least one in each experimental condition. We re-drew the figure and estimated the actual scores reported. Employing the reported agreement scores we could then calculate the range of data the second observer could have obtained on any particular session.

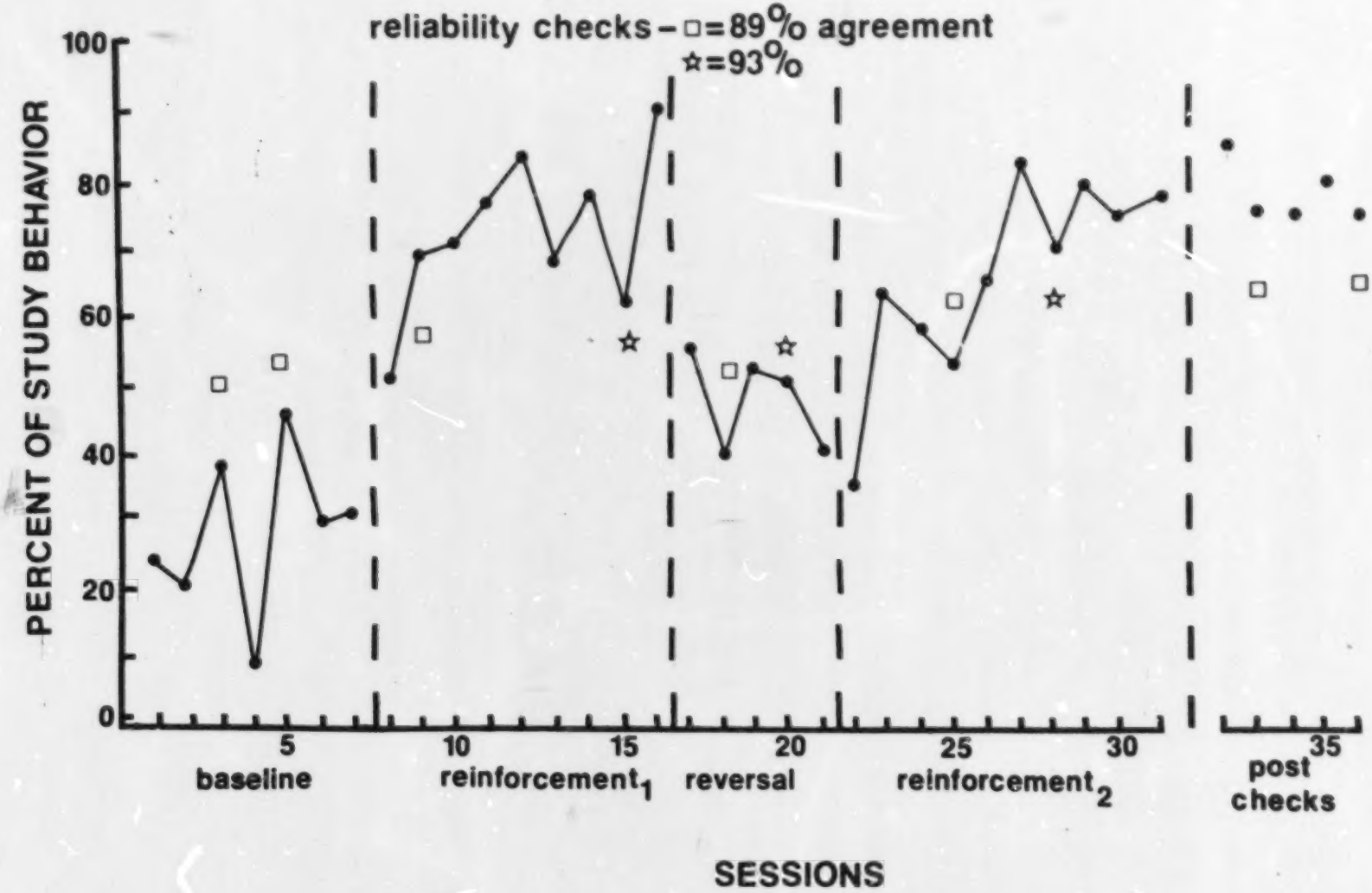
The agreement scores ranged from 89% to 93% (no mean is reported). With a 89% I-I agreement score, the second observer's data would have deviated 11 percentage points from the data reported by the primary observer. We arbitrarily selected sessions on which the reliability checks "occurred"; and we assumed two reliability checks per experimental phase, though we knew that only one was actually conducted in some phases. We then plotted fictional data that a second observer could have obtained, spacing the reliability checks a few sessions apart as would be typical in such research. We assumed that agreement was 93% on three of the ten reliability checks and 89% on the other seven checks. These fictional, but possible, data from a second observer are plotted in Figure 2 (the square data points), which is basically a redrawing of the Hall, Lund and Jackson (1968) figure. What appeared to be a clear effect of the experimental manipulation becomes highly questionable when one inspects the fictional data of the second observer.[8] Again, I-I reliability scores fail to serve an important scientific function, and here the problems of I-I scores are magnified due to the fact that errors of overestimation of behavior can sum with errors of underestimation. That is, if there is a bias toward finding an experimental effect the primary observer's data (and perhaps even the secondary observer's data) will show the behavior to be less frequent (or shorter duration) than it

insert Fig. 2 about here

really is during one condition and more frequent than it really is during another condition, and the degree to which I-I scores fail to reflect the disagreement between observers (the invalidity of I-I scores) under the one condition is, in a sense, added to their failure to reflect disagreement in the opposite direction under the other condition. Thus, in Figure 1 the invalidity of I-I scores to detect an error of measurement during baseline$_1$ is reflected in their insensitivity to the primary observer's data being lower than the real behavior, while the same invalidity during the first treatment condition is reflected in their insensitivity to the primary observer's data being higher than the real behavior. The insensitivity of I-I scores to disagreements between observers thus is doubled when I-I scores are intended or assumed to detect biases toward finding an experimental effect. Let us now consider the nature of the basic problem with I-I reliability scores.

The Basic Problem

The gross unreliability of I-I scores as an index of definition adequacy, observer competence or believability of experimental effects is a result of the fact that I-I scores are highly subject to influence by the rate (or duration) of the behavior being recorded (a problem pointed out by Bijou, Peterson, Harris, Allen, and Johnston, 1969). When a behavior is occurring so infrequently as to occupy very few intervals, the two observers are virtually certain to both record it in few intervals, even if they are employing a grossly inadequate definition of which the two have very different interpretations. This is exactly what produced agreement scores in the 90s for the observers recording "positive affect", "negative affect", and "acting silly". It is also the reason why two observers recording completely different, incompatible behaviors -- "writing" and "handraising" -- obtained agreement scores from 77% to 92%.

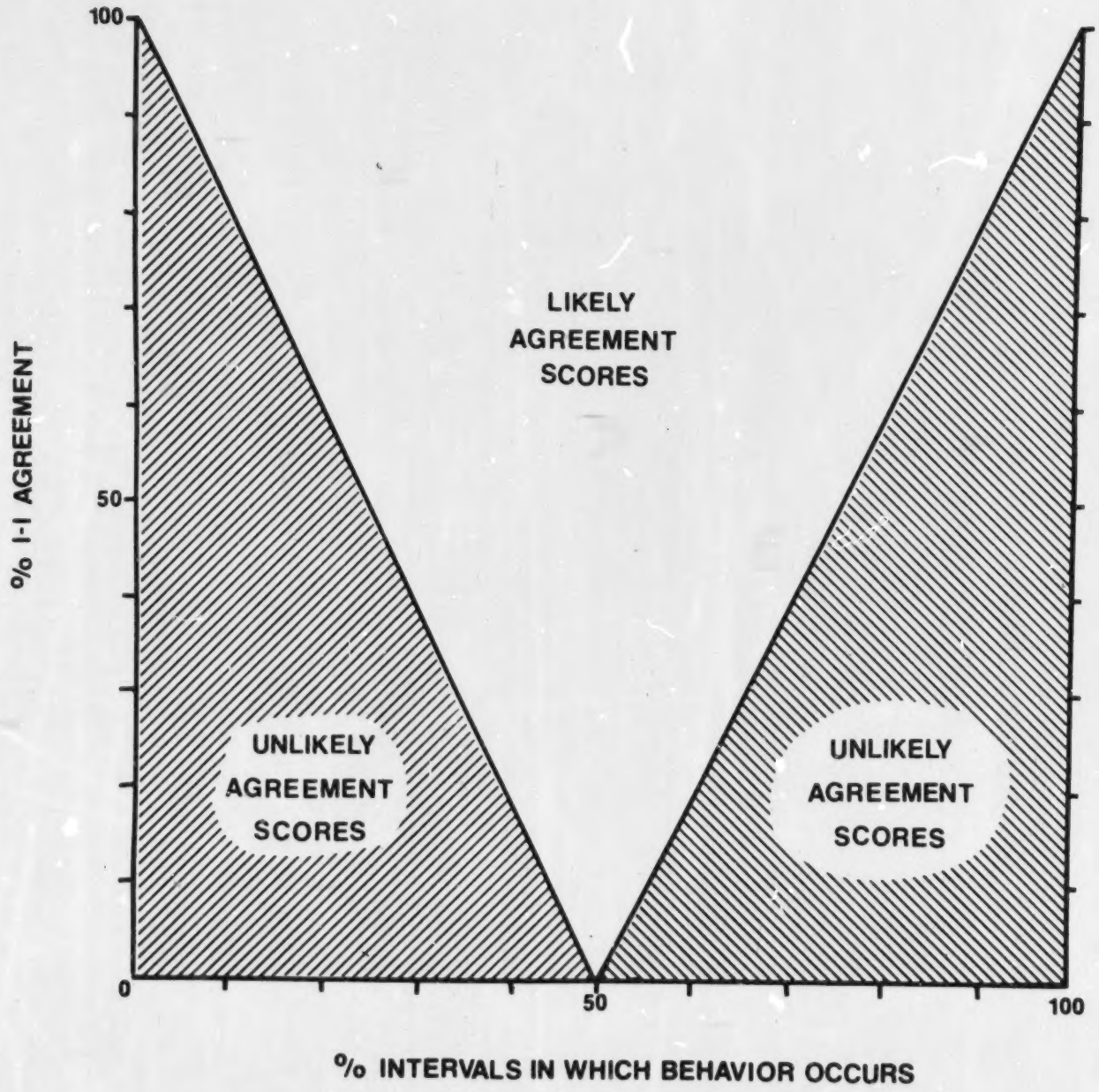

reliability checks – □=89% agreement
☆=93%
PERCENT OF STUDY BEHAVIOR
100
80
60
40
20
0
5
10
15
20
25
30
35
baseline
reinforcement1
reversal
reinforcement2
post checks
SESSIONS

Likewise, low frequency behaviors will produce high I-I agreement scores almost regardless of the incompetency of the observers. This is what occurred in the test of observer competency, where a fictitious sleeping observer agreed with an alert observer 86% to 98% on handraising and 77% to 100% on writing.

When a behavior is so frequent (or long duration) as to occur in nearly all intervals[9] the observers are likely to both record it in a high percentage of the intervals even if the definition is grossly inadequate. This occurred in the case of "neutral affect", "thinking", and "interest". Similarly, if an observer knows that a behavior will be occurring with a high frequency he can even be so incompetent as to mark it in all of the intervals before or after the session, yet the I-I scores will fail to detect this incompetence.

The relationship of I-I reliability scores to rate (or duration) of behavior is represented graphically in Figure 2. The area above the heavy line represents the range of I-I scores that are most likely, given a particular rate of behavior (actually, percent of intervals in which it is occurring). If one assumes that the two observers record the behavior in the same total number of intervals, though not necessarily in the same intervals, the heavy line separating the shaded from the unshaded portion of the figure represents the lowest <u>possible</u> I-I score for each rate of behavior. According to this analysis, the only point at which I-I reliability scores represent an index that could be very sensitive to variables other than the rate of the behavior is when the rate of the behavior is such as to occur in approximately half of the intervals. The I-I score would be of little value in serving any of the scientific functions outlined earlier if the behavior is occurring in either a small percentage or a large percentage of the intervals.

Insert 1 : 3 about here.


100
50
0
% I-I AGREEMENT
LIKELY
AGREEMENT
SCORES
UNLIKELY
AGREEMENT
SCORES
UNLIKELY
AGREEMENT
SCORES
50
100
% INTERVALS IN WHICH BEHAVIOR OCCURS

To test whether this analysis represents a real constraint on I-I agreement scores, a series of real agreement scores were obtained.[10] The data were obtained by two naive observers independently recording six student and teacher behaviors in a public school classroom. Three of the definitions were taken from Madsen, Becker and Thomas (1968) and were as follows:

On task: e.g., answers questions, listens, raises hand, works on assignment. Must include whole 10-sec interval except for Turning Around responses of less than 4-sec duration.

Turning around: Turning head or head and body to look at another person, showing objects to another child, attending to another child. Must be of 4-sec duration, or more than 90 degrees using desk as a reference. Not rated unless seated. If this response overlaps two time intervals and cannot be rated in the first because it is less than 4-sec duration, then rate in the interval in which the end of the response occurs.

Academic recognition from teacher: Calling on a child for an answer. Giving "feed-back" for academic correctness.

The other three definitions were written by the authors and were as follows:

Handraising: Any time the hand raises above the elbow in apparent attempt to gain the teacher's attention or answer a question and is free of contact with any surface such as a desk, a book or the subject's head.

Writing: Any time the subject, holding a writing utensil (pen, pencil, crayon), makes contact between the writing end of the utensil and writing material (paper, notebook, workbook, etc.). Exclude writing on the desk. Include writing on a book or something of that nature even though you think it inappropriate.

Teacher talking: Any oral sound involving the vocal cords. Include simple sounds like "oh," "huh?," "uh," and laughing aloud. It is not necessary to be able to understand the words, merely to hear the sound of the subject's voice. Whispering is excluded by the definition since it does not involve the vocal cords.

Also exclude coughing, belching, sneezing, and clearing the throat, even though they involve the vocal cords.

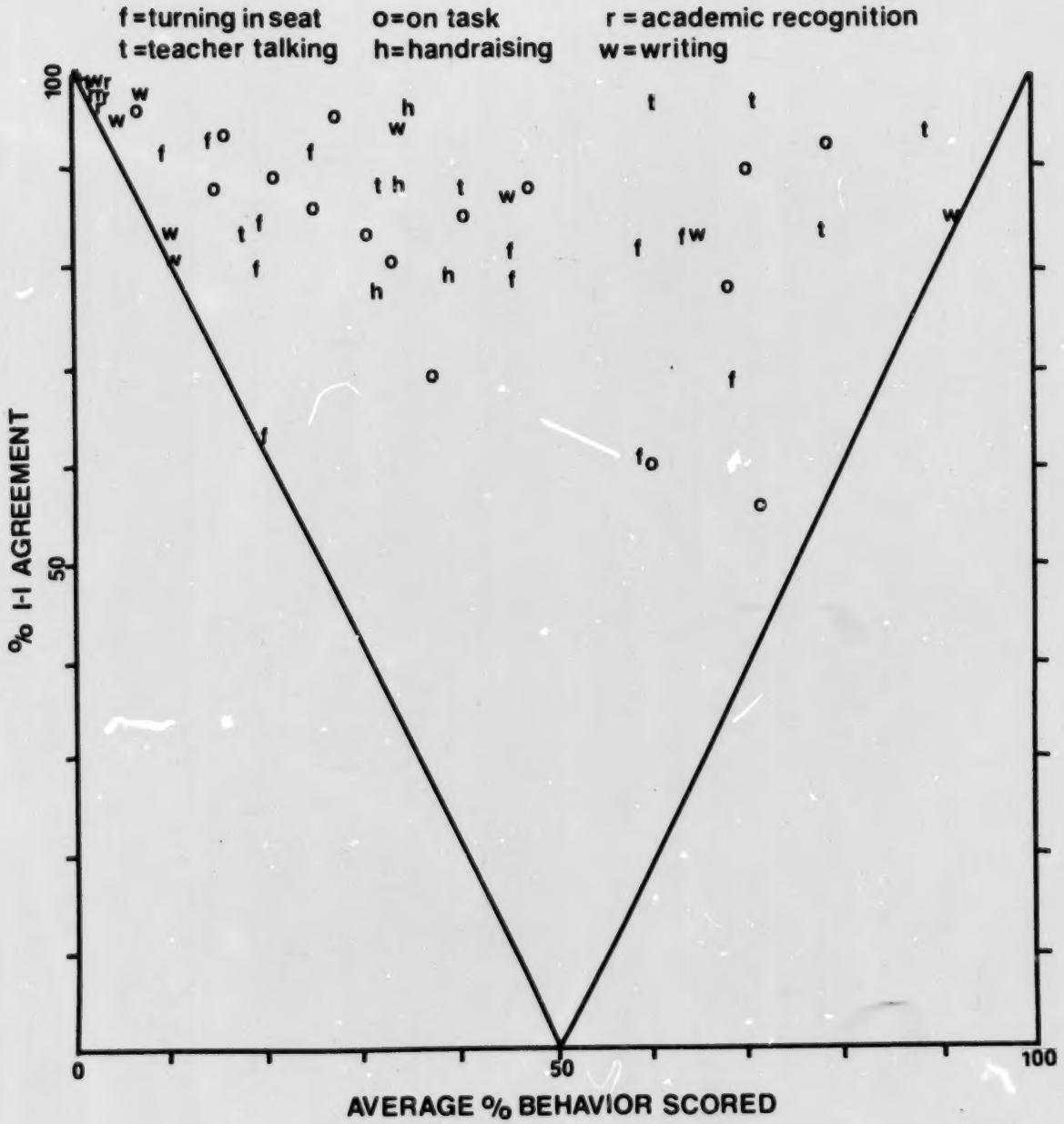
The reliability scores from 55 of the 76 pairs of observations are presented graphically in Figure 3. Each letter represents one I-I agreement score, and it is located on the ordinate to represent the value of that score and on the abscissa to represent the frequency of the behavior on that session, as determined by the mean number of intervals that the two observers recorded the behavior as occurring. The 21 scores not presented were all clustered tightly in the upper left corner of the figure, above the diagonal line, and are not presented because they would constitute a solid black area of the figure. Most of these omitted data were on the academic recognition response, and a few were on writing.

Insert Fig. 3 about here.

From the distribution of these scores it would appear that the present analysis represents a valid concern about the scientific usefulness of I-I reliability scores. All of the scores remained within the bounds suggested in Figure 2, and only toward the middle range of rates did the scores fall in the 50s, 60s, and 70s. Apparently the rate of the behavior being recorded does restrict I-I reliability scores, making them unlikely to be sensitive as an index of such variables as definition adequacy, observer competency, and observer bias.

Tentative Recommendations

Quite possibly the best ultimate solution to the problems raised in this paper will be a complex mathematical one. For example, a table or set of tables could be constructed from which the experimenter could assess the probability of an obtained difference between two observer's data, given a particular frequency of the behavior (perhaps the mean number of intervals, or perhaps two separate assessment based on the two observer's separate totals). However, the solution we wish to propose for the present is simpler. It is twofold; one aspect being aimed at solving the general problem of I-I reliability scores' being affected by the rate of behavior, and the other aspect being aimed only at the problem of the


f = turning in seat
t = teacher talking
o = on task
h = handraising
r = academic recognition
w = writing
% I-I AGREEMENT
100
50
0
50
100
AVERAGE % BEHAVIOR SCORED

believability of experimental effect.

It appears to us that the best simple solution is to shift to a different type of agreement score or a combination of two such agreement scores. One of these scores has seen occasional use in the applied behavior analysis literature (e.g., apparently Madsen, Becher and Thomas, 1968, used it). We call it scored-interval (S-I) agreement. In S-I agreement all intervals in which neither observer scored the behavior as occurring (or both observers recorded the absence of the behavior, a method used by Hall, Lund and Jackson, 1968, with part of their data, and one that may improve recording somewhat) are ignored in calculating agreement scores. Only an interval in which both observers recorded the presence of the behavior is counted as an agreement. In effect, this would remove the diagonal line from the left half of Figure 2, leaving reliability scores free to vary from 0 to 100 even at lower response frequencies. When the data on negative affect are treated this way, the agreement changes from an I-I score of 92% to an S-I score of 33% (4 agreements, 8 disagreements). Likewise, the 95% I-I agreement on positive affect becomes an S-I agreement of 0% (one observer never saw it, the other saw it 5 times); and the 99% I-I agreement on acting silly becomes an S-I agreement of 50% (both observers saw it in one interval, one saw it again in another interval.

But S-I reliability scores also have some serious limitations. First, S-I scores constitute a very stringent test of observer agreement, and new standards would clearly have to be established as to the levels of agreement our science should expect of interval data. Second, at extremely low frequencies of behavior S-I scores become highly variable. For example, if one observer recorded a response as occurring in one interval and the other recorded it as occurring in no intervals, their agreement is 0%, whereas a change of just one interval on the second observer's data sheet would make this 100%. Third, with extremely high frequencies of behavior S-I reliability scores would appear to have a liability

like that of I-I scores, they inevitably become high.

Perhaps a suggestion of Bijou et al. (1969) provides a method to counteract the above three limitations of S-I scores. These authors suggest that two reliability scores might be calculated, one for the occurrence of the behavior (S-I scores) and another for the non-occurrence of the behavior. In this second type of score an agreement is counted only when both observers recorded that the behavior did not occur. A disagreement is counted when one observer recorded the presence of the behavior and the other recorded that it was absent. Intervals in which both observers scored the behavior are ignored. This kind of agreement ratio might be called unscored-interval (U-I) reliability.

The characteristics of U-I agreement scores are such that they complement those of S-I scores. When the frequency of behavior is above 50% of the session, U-I scores can vary from 0% to 100%; but with lower frequencies of behavior U-I scores have the same lower limits as those shown on the left side of Figure 2.[11]

Although U-I scores by themselves would have the same three kinds of limitations indicated above for S-I scores, when U-I and S-I scores are presented in combination these limitations should be eliminated or greatly ameliorated. Whether the combination of U-I and S-I should be in the form of simply presenting both scores, presenting the mean of the two scores, or some other method of combination is uncertain. If both scores are presented, the researcher is still in the position of suffering from the high variability of one score or the other at extremely high or low behavior frequencies, whereas averaging the two scores reduces this problem.

To further understand properties of S-I scores, U-I scores and the mean of these two, it may be instructive to inspect these scores when derived from some of the same data on which I-I scores were reported earlier in this paper. Table 1 presents several such comparisons. As the scores in Table 1 suggest,

Insert Table 1 about here

Table 1
Illustrative Comparisons of Four Different Reliability Scores Derived from the Same Data

BEHAVIOR	INTER-OBSERVER AGREEMENT SCORES			
	I-I	S-I	U-I	X of S-I + U-I
From Adequacy-of-definition Function:				
positive affect	95%	0%	95%	47.5%
negative affect	92%	41%	92%	67%
neutral affect	100%	100%	(no interval unscored)	100% (no X possible)
thinking	83%	78%	41%	59.5%
excessive movement	57%	14%	37%	25.5%
interest	84%	80%	32%	56%
acting silly	99%	50%	99%	74.5%
From Observer-Competency Function:				
teacher talking				
Session #1	4%	0%	4%	2%
Session #2	7%	0%	7%	3.5%
Session #3	2%	0%	2%	1%
Session #4	54%	0%	54%	27%
handraising				
Session #1	88%	0%	88%	44%
Session #2	98%	0%	98%	49%
Session #3	86%	0%	86%	43%
Session #4	98%	0%	98%	49%

I-I scores are always equal to or larger than S-I scores, U-I scores or the mean of S-I and U-I.

In general, it appears that the mean of S-I and U-I is a promising statistic with which to represent the degree to which two independent observers agree on the occurrence of behavior, as measured by interval recording. However it should be noted that the mean is not free of influence by the frequency of the behavior. The nature of the relationship between the frequency of the behavior and the mean of S-I and U-I is similar to that represented in Figure 2, except that the curve's high points are 50% agreement rather than 100%. That is, if the two observers agree perfectly on the total number of intervals in which the behavior occurred, the lowest possible agreement score (mean of S-I and U-I) is 0%, and that is possible only if the observers both saw the behavior in 50% of the intervals. As the agreed-upon frequency approaches either 0% of the session or 100% of the session, the lower limit of agreement scores rises toward 50%. These lower limits appear to be much more tolerable than those on I-I scores (represented in Figure 2), and of course they decline as the two observers' agreement on total frequency decreases.

An additional comparison between the characteristics of I-I scores and those of the mean of S-I and U-I should be pointed out. It is evidenced in the lower portion of Table 1, in which agreement scores from "the sleeping-observer test" are presented. When one observer is asleep (fails to see the behavior in any interval), the lower limit of I-I scores is a straight line ranging from a low of 0% (when the behavior is seen in every interval by the awake observer) to a high of 100% (when the behavior is seen in no interval by the awake observer), while the lower limit for the mean of S-I and U-I ranges from a low of 25% (when the awake observer sees the behavior in half of the intervals) to a high of 50% (when the awake observer sees the behavior in every interval or in no interval).[12] Since agreement scores of 50% and less are unlikely to ever be

acceptable in applied behavior analysis, the lower limits of the mean of S-I and U-I seem more tolerable than those of I-I.

We wish to make two further suggestions that deal specifically with the issue of believability of the experimental effect. First, it appears essential that researchers make a practice of presenting the actual data obtained by the second observer, when reliability checks are made, rather than merely a derived agreement score.[13] The best means of presenting these data is to plot them in figures, along with the data of the primary observer, as is done with the fictional data in Figure 1 of this paper. Some researchers have begun such a practice already. It allows the reader of a research paper to more adequately interpret the agreement scores presented and to inspect the relationship between the two observers' scores across conditions in order to evaluate the possibility of observer bias (actually of differential bias between observers, since either or both could be biased to find a particular effect). For example, a reader may then note that on sessions where reliability was assessed the primary observer's data tended to show a weaker experimental effect than they showed on other sessions, as is evidenced on sessions 13 and 23 of Figure 1 (but not on other sessions where reliability was checked). Frequent occurrence of this phenomenon would suggest that the primary observer is biased to exaggerate the effect. Or a reader may note the primary phenomenon being illustrated in Figure 1, the second observer's data show considerably less experimental effect than do the primary observer's data.

Second, it seems imperative that researchers obtain frequent reliability checks, probably a minimum of two per experimental condition (thus providing an estimate of the stability of the relative biases of the observers) and a minimum of approximately one every six sessions. In addition, reliability should be assessed during sessions when the primary observer is reporting maximal experimental effect, such as sessions 13, 14, 15, 24 and 25 in Figure 1.

These precautions should provide a reasonable safeguard against the reporting of effects that are grossly distorted by observer (or similar) biases, provided the second observer does not have the same biases as the first.

Conclusion

If behavior analysts wish to continue the use of interval recording, better methods of assessing the reliability of interval data are sorely needed. Because I-I reliability scores are clearly inadequate for any of our scientific purposes, it is already likely that a significant body of applied behavior analysis has seriously misrepresented to us the relationships between certain environmental factors and certain significant human behaviors. Further, it is likely that we are making important programming decisions on the basis of such false information, decisions that affect the lives of thousands of people. Finally, it is likely that we are teaching this same false information to others, thus perpetuating and magnifying our mistakes. While a change to S-I and U-I reliability scores may not truly solve the problem, it is an easy change to make and one that appears to offer much more accurate representation of the objectivity and accuracy of interval data. When combined with additional safeguards aimed specifically at assessing the believability of the experimental effect, these reliability measures should imporve the methodology of a significant portion of applied behavior analysis in education.

FOOTNOTES

[1]Approximately 6 studies have been reported in JABA that utilized interval recording but were not conducted in educational settings. Four of them calculated inter-observer agreement by the I-I method.

[2]Some researchers do not describe their calculation of reliability scores clearly enough that a reader can determine which method was employed.

[3]The personal and social significance of these behaviors is in no way being questioned; in fact the behaviors were selected because they are significant. But anyone experienced at precise, objective measurement will quickly recognize how insufficient and overinclusive the definitions are, and how much subjective judgement will be required by the observer employing them.

[4]We wish to thank Michael Romaniuk and Larry Herman for their cooperation and assistance.

[5]It is common practice (though not universal) to actively record only the occurrence of a response, recording nothing for the non-occurrence of that response. Thus a blank data sheet might be that of an alert observer who noticed that the behavior never occurred, or it might be that of an observer who fell asleep early in the session. Although active recording of the non-occurrence of a response is probably desirable, the present analysis and argument do not hinge on the passive "recording" of non occurrences. The sleeping observer could also be simulated by arbitrarily marking a data sheet to indicate 100% non-occurrence (or any other pattern of occurrence).

[6]Had the second observer been given cues as to the expected effects, as is probably more typical in applied behavior analysis to date, the agreement scores might well have been even higher.

[7]The authors are indebted to Vance Hall, Diane Lund and Deloris Jackson, and to the Journal of Applied Behavior Analysis, for permission to use the figure presented. This analysis in no way constitutes a criticism of this particular study. Many other studies with less frequent reliability checks, lower agreement scores, or smaller-magnitude changes in behavior could be criticised more readily.

[8]The reader may note that if certain other sessions had been selected as the reliability sessions, and if it had been assumed that nine of the ten reliability checks had produced 93% agreement, the second observer's data would also show a clear experimental effect. However, it should be remembered that if the primary observer is biased, he would be most likely to show his minimal bias on days when his reliability is being checked (some relevant evidence is provided by Romanczyk, Kent, Diament, and O'Leary, 1973), and therefore selection of sessions on which the primary observer's data were nearer the median of all the data is probably justified. In addition, even if there were a real experimental effect, behavior analysts need to know the approximate magnitude of that effect, not simply that it was present; and only by selection of the most extreme data points of each condition could one force the second observer's data to show a magnitude of effect comparable to that reported by the primary observer. Other criticisms of the present method of analysis can also be raised, but the point remains that I-I scores alone cannot be confidently relied upon to assess the believability of the experimental effect.

[9]This is more likely to occur as one increases the size of the interval.

[10]We owe Robert Dobes a debt of thanks for allowing us to use some of his data.

[11]At first impression it might appear that U-I scores and S-I scores on the same data would always add up to 100%; but this proves not to be true, generally, because of the fact that any interval in which one observer scored the behavior as present and the other scored it as absent is included in the calculation of both S-I and U-I. Thus, while the numerators of the two ratios always sum to the total number of intervals in the session, the denominators add up to more than that total (except in the case where either U-I or S-I is 100%.)

[12]Actually, instead of reaching 50% agreement, the curve falls off to zero suddenly at the two extremes. This is due to the fact that an S-I score cannot be calculated (and therefore cannot be averaged with the other score) if the behavior was never recorded by either observer, and a U-I score cannot be calculated if the behavior was recorded in every interval by both observers.

[13]A fourth kind of reliability score, total-interval agreement (T-I) also serves as a safeguard against the reporting of grossly exaggerated experimental effects, but it is less effective than the simple plotting of the second observer's data. In calculating T-I agreement the researcher simply divides the total number of intervals in which one observer saw the behavior into the number in which the other observer saw the behavior, always dividing the larger into the smaller, and multiplies by 100. This provides a much better assessment of the believability of the experimental effect than do other agreement scores, because it employs the same statistic as that used in presenting the experimental effect -- the number (or percent) of intervals in which the behavior was seen. The other three reliability scores are all based on when the behavior was seen rather than how often.

REFERENCES

Bijou, S. W., Peterson, R. F., and Ault, M. A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. Journal of Applied Behavior Analysis, 1968, 1, 175-191.

Bijou, S. W., Peterson, R. F., Harris, F. R., Allen, K. E., and Johnston, M. S. Methodology for experimental studies of young children in natural settings. The Psychological Record, 1969, 19, 177-210.

Hall, R. V., Lund, D., and Jackson, D. Effects of teacher attention on study behavior. Journal of Applied Behavior Analysis, 1968, 1, 1-12.

Hart, B. M., Reynolds, N. J., Baer, D. M., Brawley, E. R., and Harris, F. R. Effect of contingent and non-contingent social reinforcement on the cooperative play of a preschool child. Journal of Applied Behavior Analysis, 1968, 1, 73-76.

Madsen, C. H., Jr., Becker, W. C., and Thomas, D. R. Rules, praise and ignoring: Elements of elementary classroom control. Journal of Applied Behavior Analysis, 1968, 1, 139-150.

Romanczyk, R. G., Kent, R. N., Diament, C., and O'Leary, K. D. Measuring the reliability of observational data: A reactive process. Journal of Applied Behavior Analysis, 1973, 6, 175-184.